Readings from

SCIENTIFIC AMERICAN

THE MOLECULES OF LIFE

W. H. Freeman and Company
New York

THE COVER

The cover displays an end-on view of the DNA double helix, the molecule that encodes genetic information and has become the emblem of molecular biology, the theme of this book. The computer-generated image offers a wide-angle look along the axis of the *B* form of the double helix. The sugar and phosphate groups comprising the backbone of one of the two strands of the molecule are diagrammed in red, the elements of the other backbone in green. The strands are linked by paired bases: a purine (*blue*) on one strand pairs with a pyrimidine (*pink*) on the other strand. The swirling cloud of dots is the solvent-accessible surface: the outermost surface of the DNA molecule, defined by individual atoms, with which other molecules interact. Arthur J. Olson of the Research Institute of Scripps Clinic generated the image. He worked with the computer-graphics language GRAMPS (which he developed with T. J. O'Donnell), a molecular-modeling package (developed with Michael L. Connolly) called GRANNY and MS, a program for calculating dot surfaces, written by Connolly.

Library of Congress Cataloging in Publication Data

Main entry under title:

The Molecules of Life.

"Readings from Scientific American."
"The eleven chapters in this book originally appeared as articles in the October 1985 issue of Scientific American."—T.p. verso.
Bibliography: p.
Includes index.
1. Molecular biology—Addresses, essays, lectures.
I. Scientific American.
QH506.M665 1986 574.8'8 85-27598
ISBN 0-7167-1792-1
ISBN 0-7167-1783-2 (pbk.)

The eleven chapters in this book originally appeared as articles in the October 1985 issue of SCIENTIFIC AMERICAN.

Printed in the United States of America

2 3 4 5 6 7 8 9 0 KP 4 3 2 1 0 8 9 8 7

CONTENTS

Proteins

Proteins are the molecules encoded by genes. The proteins in turn give rise to structure and, by virtue of their selective binding to other molecules, make genes and all the other machinery of life

by Russell F. Doolittle

If DNA is the blueprint of life, then proteins are the bricks and mortar. Indeed, they serve also as the jigs and tools needed in the assembly of a cell or an organism, and they even play the role of the builders who carry out the work of assembly. Your genes supply the information, but you *are* your proteins.

Like DNA, a protein is a linear polymer: a chain of subunits linked in a continuous sequence. In other respects, however, the two kinds of molecule are quite different. Roughly speaking, all DNA molecules are alike in overall structure, and they all have the same function (that of a genetic archive). Proteins, in contrast, fold up into a remarkable diversity of three-dimensional forms, which give them a corresponding variety of functions. They serve as structural components, as messengers and the receptors of messengers, as markers of individual identity and as weapons that attack cells bearing foreign markers. Some proteins bind to DNA and thereby regulate the expression of genes; others take part in the replication, transcription and translation of genetic information. Perhaps the most important proteins are the enzymes, the catalysts that determine the pace and the course of all biochemistry.

In the study of proteins a major aim has been to decipher their structure and so learn how they work. A complete structural analysis is a laborious undertaking, and up to now biochemists have gained a thorough understanding of only a small fraction of the known proteins. Nevertheless, some general principles have emerged; substructures that are common to diverse proteins, and that probably have similar functions in many of them, can now be recognized. Of equal interest is the question of how the thousands of proteins in a typical organism have evolved and diversified. The presence of shared substructures implies a complex evolution. It is not simply a matter of one protein's being modified and thus giving rise to another; rather, fragments of genetic information must somehow be exchanged and then expressed in many proteins.

Through all the functional diversity of proteins there runs a common thread: for the most part, proteins work by selectively binding to molecules. In the case of a structural protein the binding often links identical molecules, so that many copies of the same protein aggregate to form a larger-scale structure such as a fiber, a sheet or a tubule. Other proteins have an affinity for a molecule different from themselves. Antibodies, for example, bind to specific antigens; hemoglobin binds to oxygen in the lungs and then releases it in distant tissues; regulators of genetic expression bind to specific patterns of nucleotide bases in DNA. Receptor proteins embedded in the cell membrane recognize messenger molecules (such as hormones and neurotransmitters), which may themselves be proteins that have a specific affinity for the receptors. Virtually all the activities of proteins can be understood in terms of such selective chemical binding.

The binding of a protein to the molecule it recognizes is not fixed or permanent. It is governed by a dynamic equilibrium, in which molecules are continually being bound and released. At any instant the percentage of bound molecules depends on the relative amounts of the two substances present and on the strength of the association between them. The binding strength depends in turn on how well the molecules fit together geometrically and on specific local interactions, such as electrostatic attraction or repulsion between charged regions.

Enzymes, in this respect, are much like other proteins. An enzyme recognizes a specific molecule (called the substrate) and binds to it in dynamic equilibrium; what distinguishes an enzyme is that it can bring about some chemical change in the bound substrate. The change generally entails the forming or breaking of a covalent chemical bond: the substrate may be split into two pieces, a chemical group may be added or the pattern of the bonds in the substrate may simply be rearranged.

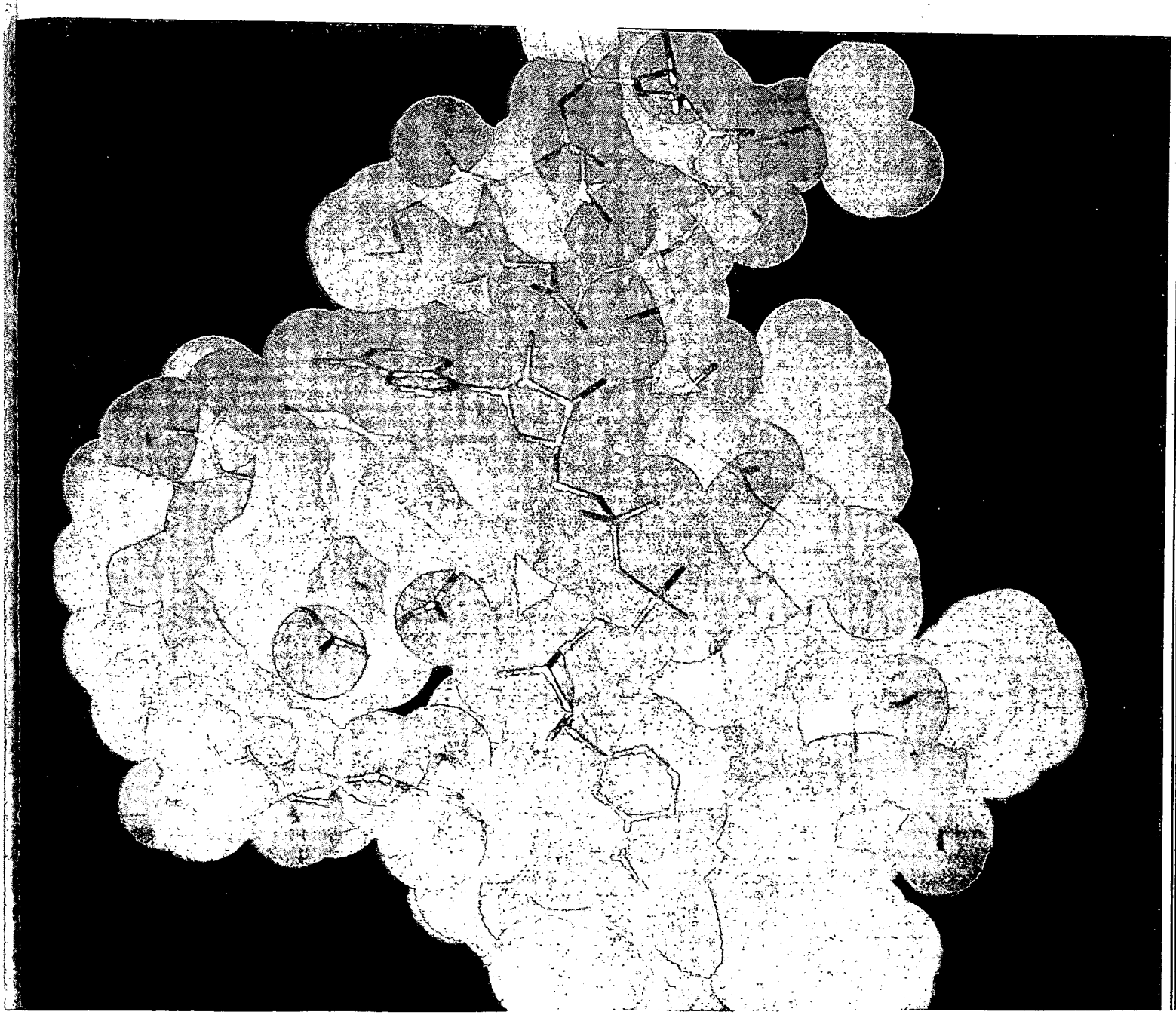
The mechanism of enzyme action can be viewed as having three stages. First the enzyme binds to the substrate, then the chemical reaction takes place and finally the altered substrate is released. All three steps are reversible. If an enzyme binds to molecule X and

PROTEIN BINDING SITE is emblematic of the principal mechanism by which proteins do the work of biochemistry: by forming a close but generally short-lived association with another molecule. The protein is alcohol dehydrogenase, an enzyme in the liver that converts ethyl alcohol into acetaldehyde. Carbon atoms in the protein structure are white, oxygen atoms red and nitrogen atoms blue. The atoms shown in purple make up a molecule of nicotinamide adenine dinucleotide (NAD), a coenzyme that takes part in the catalyzed reaction by receiving a hydrogen ion removed from an alcohol molecule. (The alcohol is bound to a site elsewhere on the protein.) The NAD molecule fits precisely into a cleft on the protein surface and is held there by electrostatic attraction. Many proteins that bind to NAD and related coenzymes include a domain of similar structure. It is called the mononucleotide fold and may be one of the most ancient structural units in the evolution of proteins. This computer-generated image and the ones on pages 42 and 43 were made by Jane M. Burridge of the U.K. Scientific Centre of the International Business Machines Corporation.

converts it into molecule *Y*, the same enzyme can also bind to *Y* and change it back into *X*. Indeed, there are many possible reaction paths. A molecule of either *X* or *Y* could be bound but released before any change took place, or a molecule of *X* could be converted into *Y* and then changed back into *X* before it was released, and so on.

It should be emphasized that the enzyme itself does not determine the direction of the reaction. The proportion of *X* and *Y* at equilibrium depends on thermodynamic considerations; the favored proportion is the one that minimizes the quantity called free energy. (Roughly speaking, the free energy of a system is equal to its energy minus its entropy, or disorder.) The enzyme merely hastens the attainment of equilibrium. Nevertheless, an enzyme can effectively control the course of a biochemical process. In the absence of an enzyme most biochemical reactions are extremely sluggish; the appropriate enzyme can speed them up by a factor of a million or more. Although the enzyme has no influence on whether more *X* is converted into *Y* or vice versa, it determines whether or not the conversion takes place at all.

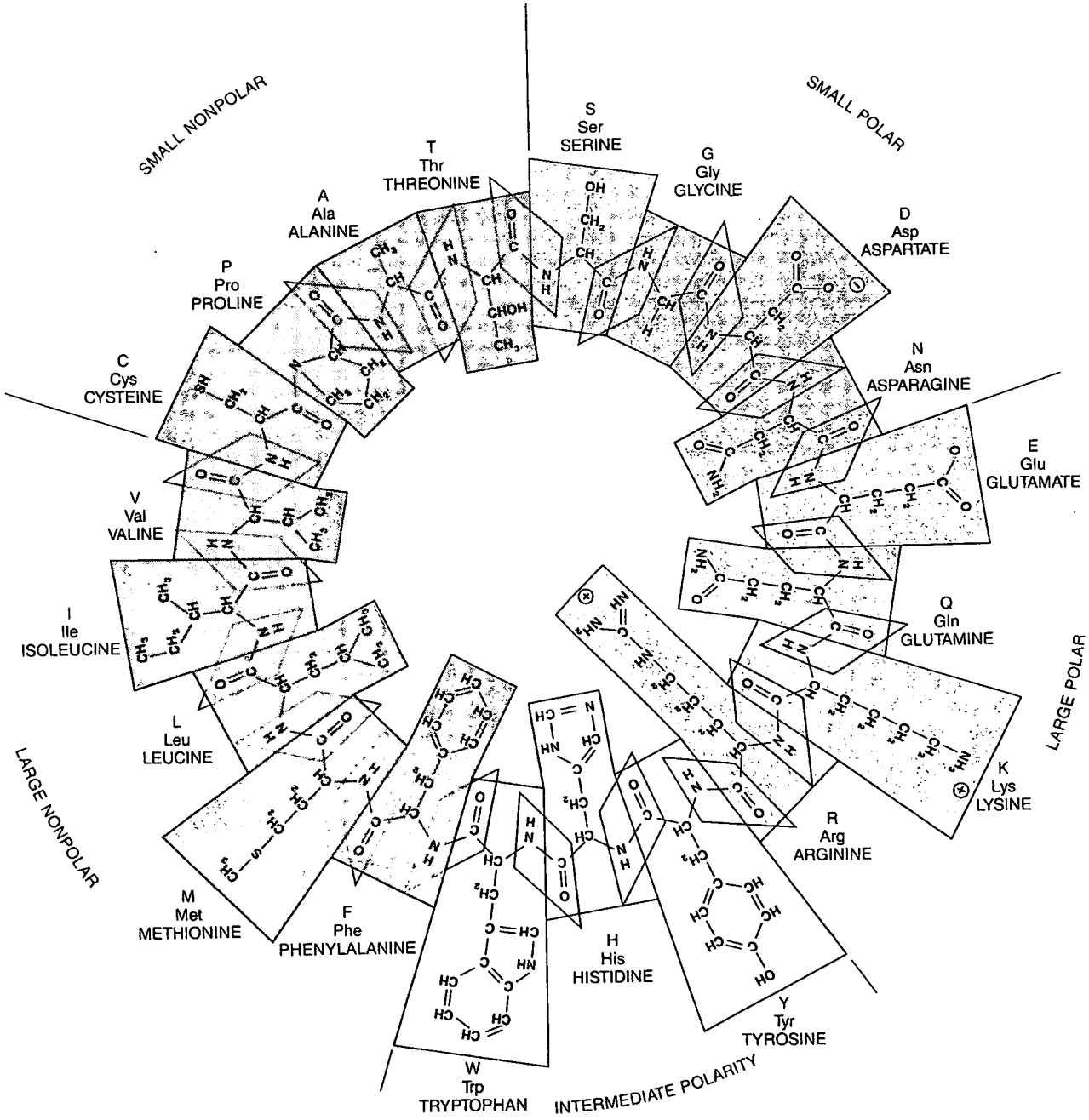
An enzyme speeds a reaction by lowering an energy barrier. Even when a reaction is thermodynamically favorable—when the products have a lower free energy than the reactants—there may be an intermediate state with a higher free energy. The enzyme tends to smooth this hump in the reaction path. The mechanism varies from case to case. Some enzymes merely provide an environment different from that of the aqueous medium, or they bring the reactants into close contact. Other enzymes take a more active role by adding or subtracting a proton, by straining bonds in the substrate molecule or even by forming transient covalent bonds between the substrate and some part of the enzyme itself. Certain enzymes are helped by the accessory molecules called coenzymes. The coenzyme binds to a specific site on the protein and provides chemical functions that are not available in the enzyme itself.



More than 2,000 enzymes have been identified on the basis of the chemical reactions they catalyze. All these proteins must be structurally distinct; in other words, proteins must come in at least 2,000 forms capable of recognizing specific molecules. How are these diverse structures generated? The "alphabet" from which proteins are made consists of the 20 amino acids that can be specified in the genetic code; every protein is a sequence of amino acids drawn from this alphabet. The physical and chemical properties of a protein molecule depend on how the chain of amino acids folds up in three-dimensional space.

All the information needed to define the three-dimensional structure of a protein is inherent in the amino acid sequence. As the chain is constructed on the ribosome, it folds up in the way that minimizes the free energy; in other words, the chain assumes its "most comfortable" configuration. In principle, if one knew all the forces acting on the thousands of atoms in the protein and on the surrounding solvent molecules, one could predict the three-dimensional structure from knowledge of the sequence alone. Such a calculation is not now feasible.

The 20 amino acids are all built on a



TWENTY AMINO ACIDS specified in the genetic code are the basic components of all proteins. Here the amino acids are shown joined head to tail to form a ring (which is not the structure of any real protein); their three-letter and one-letter abbreviations are indicated. The arrangement places amino acids that have similar chemical properties near one another in the ring. An approximate classification in five groups is based on the size of the amino acid's side chain and on the degree to which it is polarized. (A polar molecule has separated regions of positive and negative electric charge.) These factors have a major influence on the folding of a protein. In the evolution of a protein a mutant form is more likely to be accepted if an amino acid is replaced by one that has similar properties—by one found nearby in the ring. The ring is similar to one proposed by Rosemarie M. Swanson of Texas A&M University.

common foundation. They have an amino group (NH_2) at one end and a carboxylic acid group (COOH) at the other end; both groups are attached to a central carbon atom called the alpha carbon. Also attached to the alpha carbon are a hydrogen atom and a fourth group called the side chain. It is only in the nature of the side chain that the amino acids differ from one another.

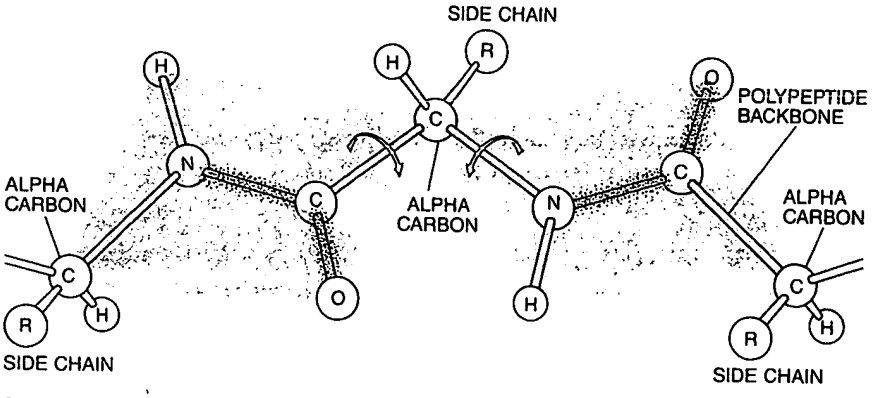
The backbone of the protein is built by linking amino acids head to tail: the amino group of one unit is joined to the carboxyl group of the next. The fusion is accomplished by removing a molecule of water, leaving the structure –CO–NH–. The carbon-nitrogen linkage created in this way is called a peptide bond, and the protein chain is referred to as a polypeptide.

The properties of the peptide bond impose certain constraints on the folding of the protein. Electrons are shared among the oxygen, carbon and nitrogen atoms in a way that gives the bond torsional stiffness; it resists rotation about its axis. As a result each peptide-bond unit lies in a plane, and the chain must fold almost entirely through rotations of the alpha-carbon bonds. The polypeptide backbone is not so much a flexible string of beads as it is an articulated chain of flat plates.

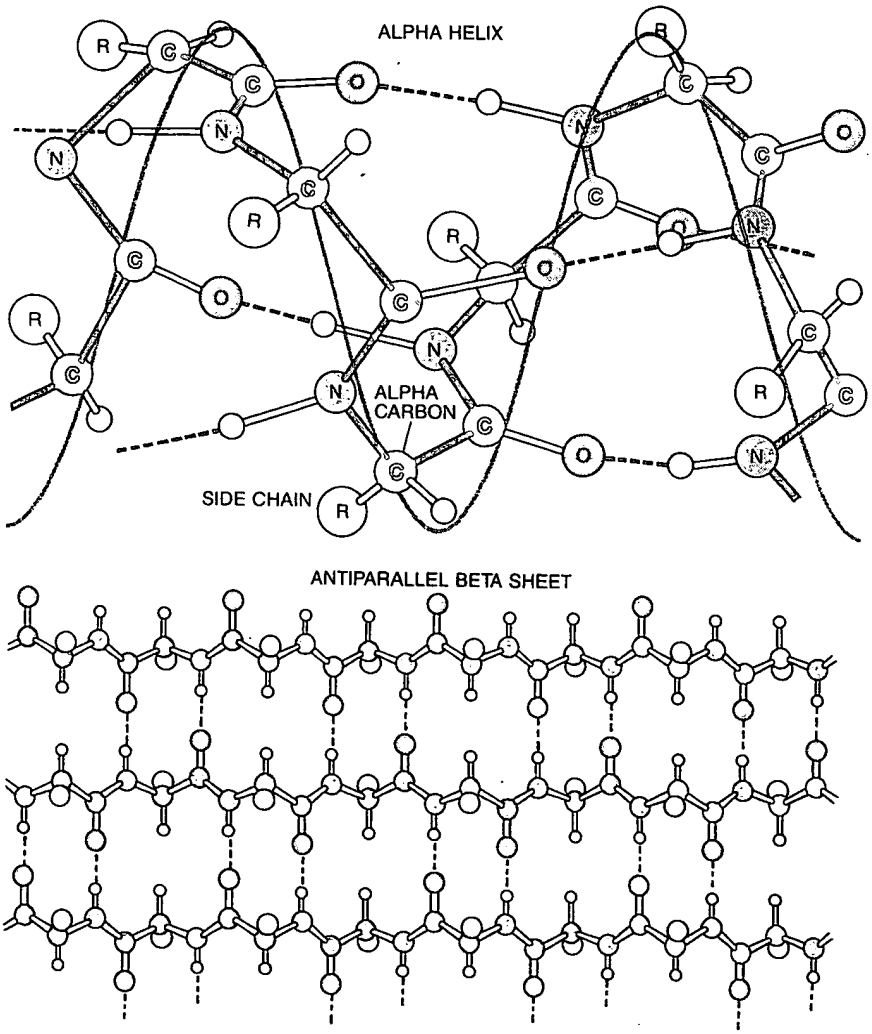
The main influence on protein folding comes from the properties of the side chains. Interactions of one side chain with another and with molecules in the medium can force the polypeptide to fold up into a compact globule with a specific, stable shape.

Some of the amino acids are polar molecules: although they are electrically neutral overall, they have localized concentrations of positive and negative charge. The polarization results from the presence of oxygen or nitrogen atoms, which have a strong affinity for electrons. A few of the amino acids not only are polar but also carry a net electric charge; in other words, they are ionized under physiological conditions. Other side chains (generally those made up exclusively of carbon and hydrogen) are nonpolar. There is a strong tendency for the polar side chains to seek a polar environment and for the nonpolar ones to be segregated in nonpolar areas. Water, the medium in which most proteins are immersed, is a strongly polar substance. When a polar or charged side chain projects into the aqueous environment, the water molecules assume an orderly arrangement. A nonpolar side chain in water disrupts this alignment of charges.

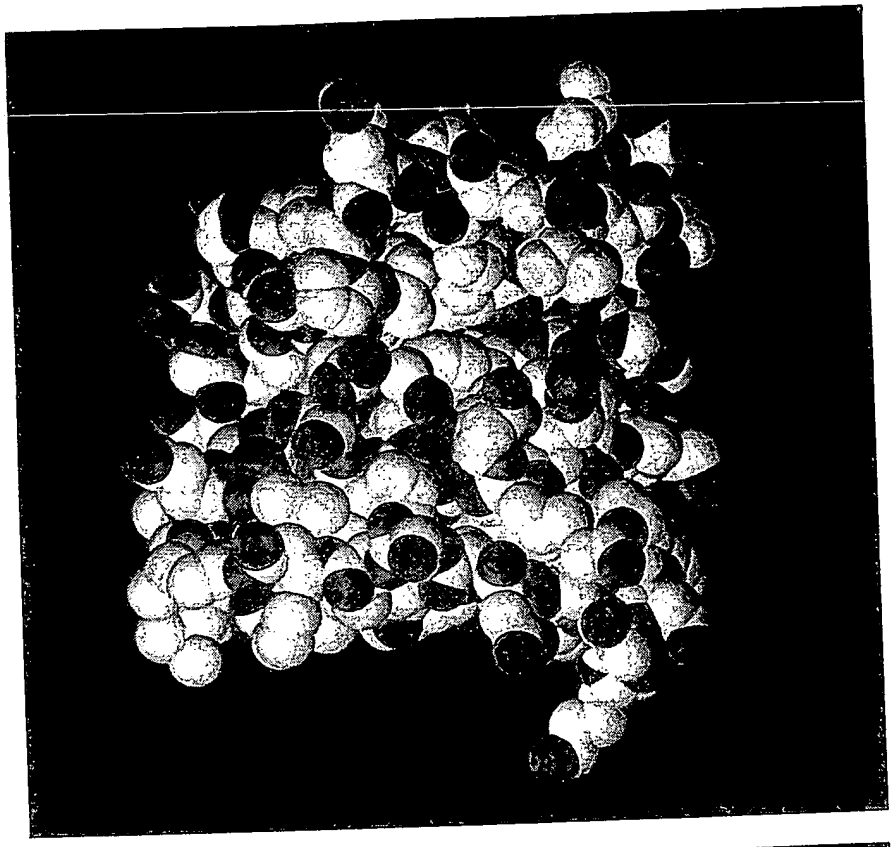
The chief consequence of these interactions is that a protein chain tends to fold so that polar side chains are on

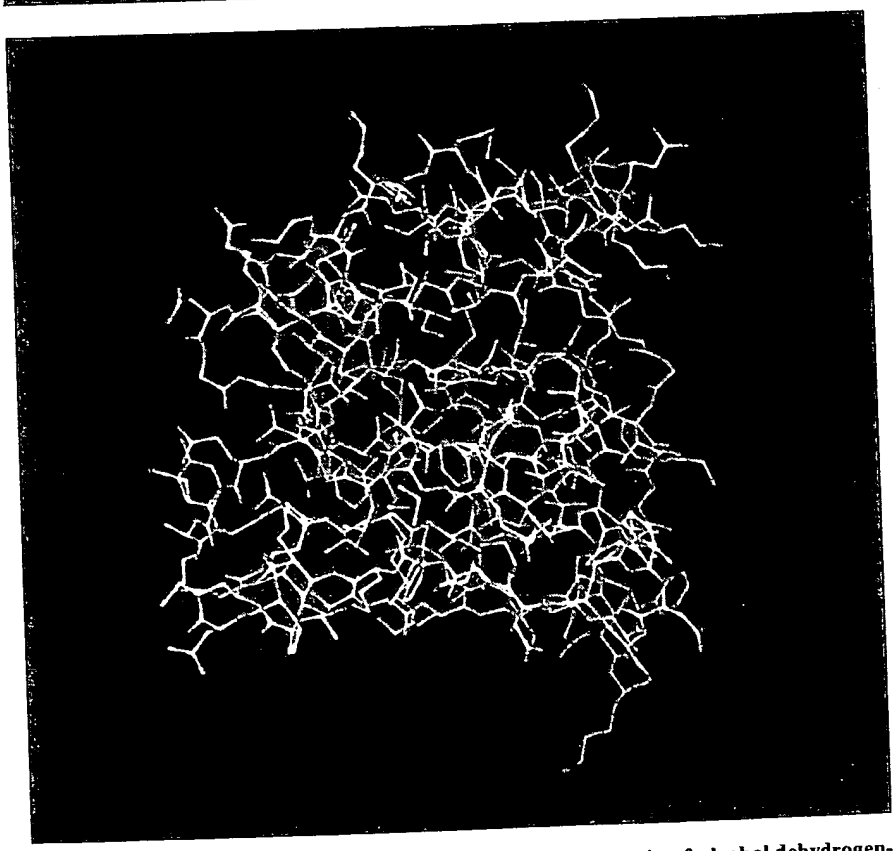


STRUCTURE OF AMINO ACIDS constrains protein folding. In an isolated amino acid four chemical groups are attached to the central, or alpha, carbon atom: an amino group (NH_3), a carboxylic acid group (COOH), a hydrogen atom and a side chain (designated *R*). The 20 amino acids differ only in the identity of their side chains. In a protein the amino group of one amino acid is linked to the carboxyl group of another (with the loss of a water molecule), forming a peptide bond. The sharing of electrons among nitrogen, carbon and oxygen atoms makes the bond resistant to twisting, so that each amino acid unit is a rigid plane. The protein can fold only by means of rotations about the bonds to the alpha carbon.



ALPHA HELIX AND BETA SHEET are common structural units of protein molecules. The sequence of amino acids in a protein is called its primary structure; as the chain is synthesized, regions of it fold spontaneously into alpha helixes and beta sheets, which constitute the secondary structure; the helixes and sheets are assembled in turn to create the tertiary structure. Both the alpha helix and the beta sheet are stabilized by hydrogen bonds (*broken colored lines*), in which a hydrogen serves as a bridge between oxygen or nitrogen atoms.





OVERALL CONFORMATION of the coenzyme-binding domain of alcohol dehydrogenase is shown in two graphic representations. The space-filling model (*upper image*) emphasizes the surface texture of the molecule; the skeletal diagram (*lower image*) reveals the internal structure. In these views the complete polypeptide backbone and all the amino acid side chains are shown, but hydrogen atoms are omitted. The folding of the chain appears to be a random jumble but is actually quite specific: every molecule of alcohol dehydrogenase folds in exactly the same way. The domain, about half of the molecule, is seen from another point of view in the illustration on page 39. Here the NAD-binding site is at bottom.

the exposed surface and nonpolar ones are inside. An exception to this rule is found in proteins embedded in cell membranes. The membrane is made up of fatty, nonpolar molecules, and the segment of the protein that passes through it likewise consists mainly of nonpolar amino acids. They anchor the protein in the membrane.

The electrostatic attraction between a polar side chain and water is a form of hydrogen bonding, in which a hydrogen atom acts as a bridge between charged oxygen or nitrogen atoms. Hydrogen bonding between one atom and another within the protein itself also helps to stabilize the structure.

Hydrogen bonds are weaker than the covalent bonds of the polypeptide backbone. Moreover, the atoms in a protein that are hydrogen-bonded to one another could as easily be hydrogen-bonded to water; the energy difference between the two configurations is small. Because many hydrogen bonds can form simultaneously as the protein folds, however, they contribute greatly to the stability of the structure.

Still another form of bonding can cross-link regions of the molecule. The amino acid cysteine has a sulfhydryl (SH) group at the end of its side chain. If the protein includes two cysteine units, they can combine to form a covalent disulfide bond (–S–S–). Such cross-links are much stronger than hydrogen bonds.

The amino acid sequence of a protein is called its primary structure. The complete three-dimensional conformation of a single polypeptide strand is referred to as the tertiary structure. As these terms suggest, there is an intermediate level of organization called the secondary structure. It describes the local folding of the chain in terms of structural units that appear in almost all proteins.

Some 35 years ago Linus Pauling showed that the protein backbone can be coiled into a tight helix stabilized by numerous hydrogen bonds; he called the structure the alpha helix. The helix makes one turn for every 3.6 amino acids, and hydrogen bonds form between amino acids four units apart. The bonds do not involve the side chains but rather extend from the NH group of one peptide unit to the CO group of another; for this reason the stability of the helix is not strongly dependent on the identity of the side chains, and many different sequences of amino acids can spontaneously assume the form of an alpha helix.

At about the same time, Pauling proposed a second stable configuration he designated the beta sheet. In this case lengths of polypeptide chain

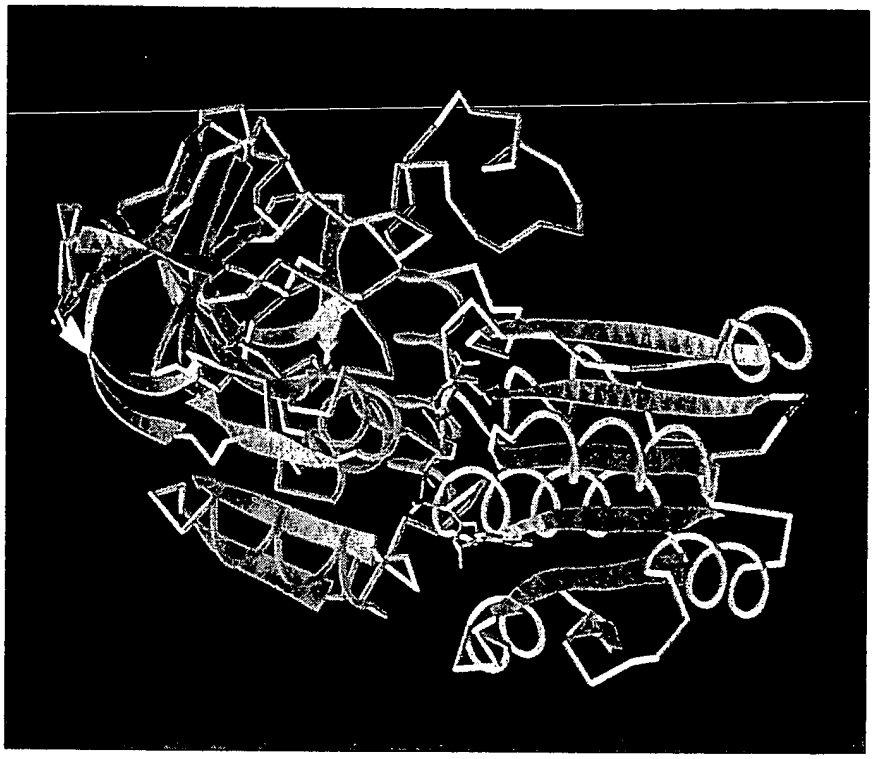
lie next to one another and run either parallel or antiparallel, with hydrogen bonds connecting the adjacent strands. Again the bonds join the NH and CO groups of the backbone.

Some proteins are composed mostly of alpha helix and others are predominantly beta sheet. In a typical globular protein the interior is a bundle of beta strands running back and forth diametrically and the surface is covered with alpha helixes. The exterior helixes generally show a characteristic periodicity in amino acid sequence. Nonpolar side chains appear at every third or fourth position and are directed toward the interior of the molecule; the rest of the side chains, which are exposed to the aqueous environment, tend to be polar.

In recent years still another intermediate level of protein structure has been perceived. For example, a structural element present in numerous proteins consists of two beta strands connected by a segment of alpha helix. The three pieces nestle together comfortably when they are arranged at particular angles. A structural feature of this kind, which typically encompasses from 30 to 150 amino acids, is called a domain. It can be considered a single unit because its conformation is determined almost entirely by its own amino acid sequence. The beta-alpha-beta domain is of particular importance because when two such domains lie next to each other, the crevice they form often serves as a binding site.

A typical globular protein includes about 350 amino acids, which could fold in innumerable ways. The hierarchy of larger-scale structures brings a measure of order. Local interactions between nearby amino acids give rise to alpha helixes, beta sheets or other forms of secondary structure. These subassemblies, acting as more or less coherent units, organize themselves into domains. The geometric arrangement of the domains constitutes the tertiary structure. The presence of the same secondary structures and domains in many dissimilar proteins argues that they are not mere artificial abstractions introduced by the biochemist; on the contrary, they seem to be fundamental units in the evolution and diversification of proteins.

Many proteins have a level of organization beyond the tertiary structure. They are composed of multiple polypeptide strands held together by a variety of weak bonds and sometimes further cemented by disulfide linkages. Some proteins also have nonpeptide components. Metal ions, for example, are essential to the activity of certain enzymes, and a structure called the porphyrin ring is found in hemoglobin, chlorophyll and a number of other proteins. Many proteins are also "decorated" on their surface with chains of sugar molecules. These additional features of protein structure are elaborations of the molecule added after the polypeptides are synthesized.



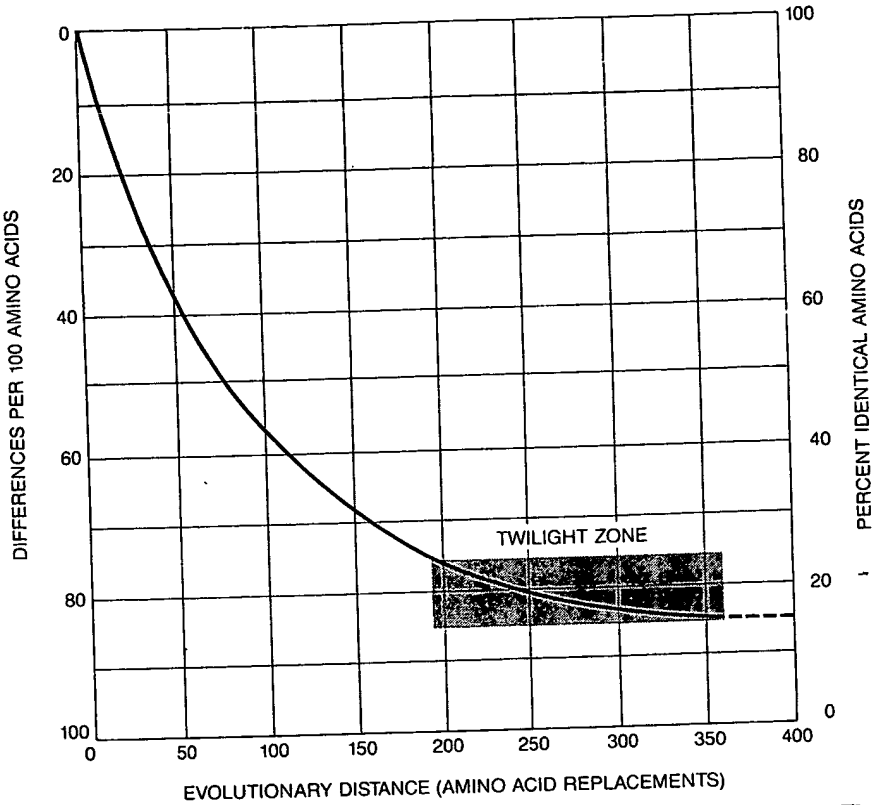
SECONDARY STRUCTURE of alcohol dehydrogenase consists of numerous alpha helixes and beta sheets connected by short lengths of "random" structure. The NAD-binding domain is in green and yellow; the catalytic domain, which binds to an alcohol molecule, is in blue. A bound NAD molecule is shown in purple near the junction of the protein domains.

In a way it is remarkable that any protein consistently assumes a single, well-defined conformation. The folded state does have a lower free energy than any alternative configuration, but the difference is small. In an alpha helix hydrogen bonding between peptide units reduces the energy, but if the helix were unraveled, the same sites would form hydrogen bonds with water. Furthermore, because a helix is an ordered structure, it has a low entropy, which tends to increase the free energy. It is worth noting that not all polypeptides have a stable folded pattern. Artificially constructed random sequences of amino acids are generally loose, flexible coils that continually shift from one structure to another. The proteins found in biological systems appear to be a subset of polypeptides selected for their stability of structure.

How is the structure of proteins discovered? Of all the methods employed by the protein chemist, the most revealing has been X-ray crystallography. The basic idea is to form a diffraction pattern by passing X rays through a crystallized specimen of the protein. Because of the periodic structure of the crystal, the pattern is essentially the same as the one that would be generated if a single molecule could be examined. From the diffraction pattern one constructs a map showing the density of electrons in the protein, and from the map the path of the backbone and the positions of the side chains can be inferred.

X-ray crystallography provides a three-dimensional view and shows a protein in atomic detail. It is through such studies that the main themes of protein structure have been elucidated: that the interior is filled with nonpolar side chains, that the alpha helix and beta sheet are more than hypothetical structures, that most proteins are compact globules with dimpled surfaces, and much more. Crystallographers also confirmed the existence of domains and discovered common patterns among them.

Ideally the three-dimensional structure of all proteins would be studied by X-ray crystallography, but that is not feasible. Crystallizing a protein in the first place often calls for a good deal of



EVOLUTION OF PROTEINS can be traced by comparing amino acid sequences. The number of amino acid positions at which two proteins differ is a measure of their evolutionary distance, but the relation is not a simple one of proportionality. A single site may undergo repeated mutations, so that the actual number of amino acid replacements is generally greater than the number of observed differences. Furthermore, in matching two sequences allowance must be made for insertions and deletions as well as substitutions. As a result, when proteins are identical at fewer than about 15 percent of their positions, common ancestry cannot be distinguished from chance coincidence. Many of the most interesting evolutionary relations lie in the "twilight zone" between 15 and 25 percent identity.

chemical wizardry, and the subsequent analysis of diffraction patterns is arduous. It took 23 years to map the structure of hemoglobin. Up to now the three-dimensional structures of only about 100 proteins have been solved.

There was a time, in the 1950's, when merely determining the amino acid sequence of a protein was also a difficult and laborious procedure, even for a very small protein. First the total composition of the protein was found by breaking all the peptide bonds in a sample of the material and measuring the amount of each amino acid present. Other samples were only partially digested, leaving small fragments whose amino acid content could then be analyzed in turn. A special chemical trick revealed which amino acid was at the amino end of each fragment. Having gathered information on many overlapping fragments, the biochemist could attempt to solve the elaborate puzzle of how the pieces fit together.

In the 1960's the technology of sequence analysis improved dramatically, and by 1970 the procedure had been automated. Amino acids were removed one at a time from the amino end of the chain and identified. A limit remained, however, on the maximum length of chain that could be handled, and so large proteins still had to be broken down into fragments.

In the past few years an indirect method of sequence analysis has all but supplanted the traditional techniques of protein chemistry. The key to the new method is that the nucleotide sequence of a DNA molecule is much easier to determine than the amino acid sequence of a protein. If one has a length of DNA that is known to encode the structure of the protein, it is a simple matter to sequence the DNA and translate each three-base codon into the corresponding amino acid.

The one difficult operation is finding the DNA that encodes the protein. In one strategy the first step is to analyze about 25 amino acids at the amino end of the protein. An appropriate segment of from five to seven amino acids within this range is then "back-translated" into a nucleotide sequence. This process is not without ambiguity: although every codon specifies exactly one amino acid, most of the amino acids can be specified by more than one codon. The key is to choose a sequence that has as little ambiguity as possible and then to produce a DNA molecule for each possible back translation. If the amino acids include a histidine unit, for example, DNA is made with both *CAT* and *CAC* codons at the appropriate position; these are the two codons that specify histidine.

The back-translated DNA serves as a probe to find complementary DNA sequences. It is labeled with a radioactive isotope of phosphorus and allowed to hybridize with the DNA in a library of cloned gene fragments. The clones with a matching sequence are readily identified by the presence of the radioactive phosphorus; they are isolated, cultured in quantity and then sequenced. The approach may seem roundabout, but it is simpler and more accurate than direct chemical analysis of protein fragments. Some 4,000 polypeptide sequences are now known.

The geneticist Theodosius Dobzhansky once wrote: "Nothing in biology makes sense except in the light of evolution." The same is true of protein structure: it makes sense only in terms of protein evolution. Just as all living organisms surely trace their lineage to a few progenitors, the great majority of proteins must be descended from a very small number of archetypes.

Evidence supporting this assertion comes from many quarters, and I shall defend it only briefly. The most straightforward argument is the manifest difficulty of "inventing" a protein de novo. As pointed out above, most random polypeptides do not even fold, much less exhibit a biological function; a new protein is far more likely to arise from modification of an existing one. There is abundant evidence of this process in specific amino acid sequences that are encoded by more than one segment of DNA in a given genome. Moreover, in proteins that fold to form localized domains crystallographers consistently find the same patterns in varied settings; once a substructure has proved useful, it seems to be called on repeatedly.

The primary mechanism of protein evolution is gene duplication, in which a cell comes to include two copies (or more) of a single gene. One copy retains its original function, so that the organism's viability is not compromised by the lack of an essential protein. The redundant copy is therefore free to mutate without constraint from natural selection. Most mutations ge

erate a nonfunctional protein, but an occasional advantageous change can create either an improved version of the original protein or a protein with an entirely new function.

There are two aspects to the study of protein evolution, which must be carefully distinguished. One can examine the "same" protein in various species, observing how the structure has changed over the course of biological time. For example, the amino acid sequence of cytochrome *c*, a protein that transfers electrons in metabolism, has been determined for more than 80 species, from bacteria to man. One product of such studies is a taxonomy of the organisms based on the relations of their proteins. The other approach is to compare the structures of various proteins within a single species. From this endeavor one can construct the family tree of the proteins themselves.

Comparisons from species to species offer considerable insight into protein chemistry. Between closely related organisms the commonest changes substitute one amino acid for another with similar properties, so that the overall structure of the molecule is not disrupted. As the evolutionary distance between the species increases, the sequences diverge. Ultimately the consanguinity of the sequences may be undetectable, even though the two proteins are unmistakably alike in tertiary structure. What this means is that completely different amino acid sequences can fold into the same shape.

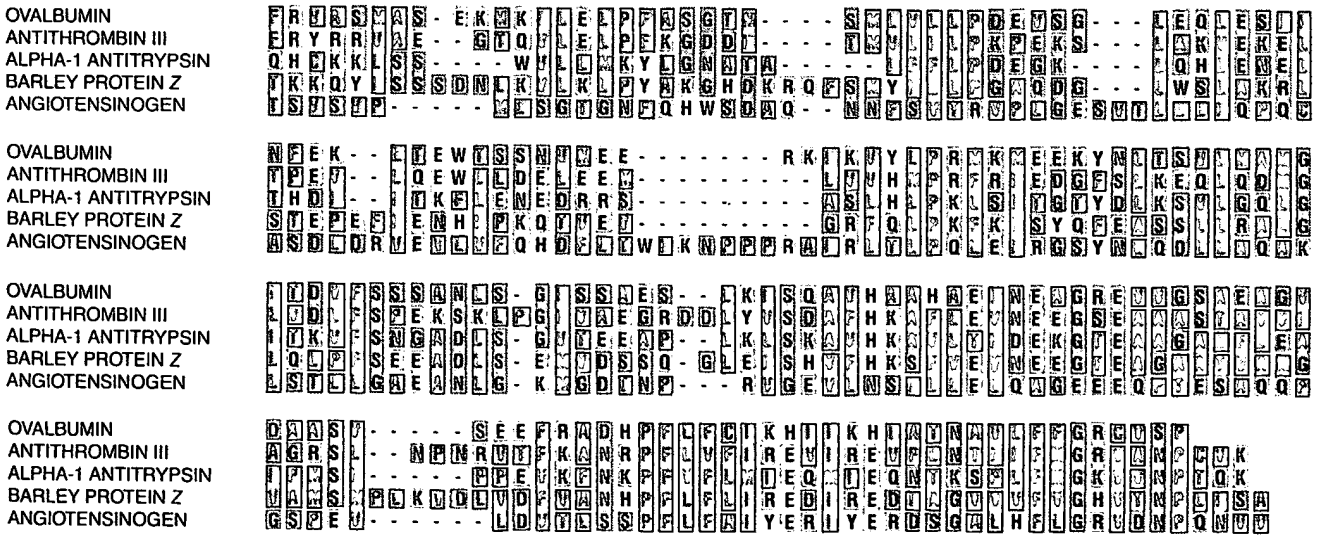
In comparing different proteins within a single species it soon becomes obvious there are broad families of related molecules. The half-dozen polypeptides that make up various forms of hemoglobin, for example, and the single polypeptide of myoglobin all share clear similarities. They are not only analogous (meaning they are similar in function) but also homologous (meaning they derive from a common ancestor). Among the enzymes it is not surprising that those catalyzing similar reactions often have homologous sequences. Glutathione reductase and lipoamide reductase provide an illustrative example. Both enzymes catalyze the transfer of hydrogen ions to sulfur-bearing compounds; they are identical at more than 40 percent of their amino acid positions. A similar degree of homology is evident between chymotrypsinogen and trypsinogen and between ornithine transcarbamylase and aspartate transcarbamylase.

As the kinship between proteins grows more remote, sequence homology becomes harder to detect. Worse, the arithmetic of sequence comparison is such that unrelated sequences may appear tantalizingly similar. Offhand, one might expect two randomly chosen polypeptides to be identical at about 5 percent of their amino acid positions; after all, there are 20 amino acids. If the comparison could be made by simply writing down the sequences one above the other and then ticking off the matches, the 5 percent limit would apply, but in reality a more sophisticated method is needed.

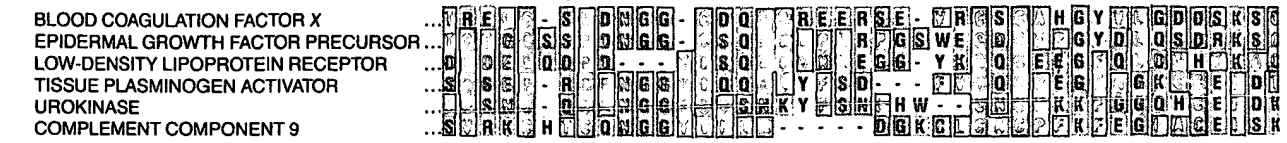
A protein can be altered not only by the substitution of one amino acid for another but also by the deletion or insertion of amino acids. Suppose two proteins are identical except that one has lost its first amino acid; if no allowance were made for this deletion, the proteins would appear to be unrelated. On the other hand, if unlimited gaps and insertions were allowed, any two proteins could be forced to match arbitrarily well. In practice the sequence comparison is done with a computer program that rewards matches between identical or similar amino acids and imposes penalties for gaps and insertions. Even so, it is virtually impossible to distinguish between chance similarity and common ancestry when the number of identical positions falls below about 15 percent.

In tracing the genealogy of proteins the relations of greatest interest are those between sequences that (after adjustment for gaps and insertions) are between 15 and 25 percent identical. This "twilight zone" is where one must look for the roots of the protein family tree, to find molecules that diverged early in the course of their evolution.

In the early 1960's it became clear that a repository of amino acid sequences would facilitate studies of protein evolution, and in 1965 Richard Eck and Margaret O. Dayhoff issued the first volume of the *Atlas for Protein*



MOLECULAR PALEONTOLOGY reveals a pattern of common ancestry for five proteins from diverse species. Each protein is represented by a sequence of the one-letter abbreviations for amino acids given in the illustration on page 40; the colors relate amino acids with similar properties. Dashes indicate gaps or insertions. Cysteine units, which can form cross-links that stabilize the folded structure, are marked by boxes. Ovalbumin is an abundant protein in egg white; antithrombin III and alpha-1 antitrypsin are found in blood plasma; barley protein Z was recently discovered in barley seeds, and angiotensinogen is the precursor of a small protein that regulates blood pressure. Both antithrombin III and alpha-1 antitrypsin are known to act as inhibitors of proteases (enzymes that cut protein chains). The functions of the other proteins had not been known, but now it seems they too may be protease inhibitors.



COMMON SEQUENCE embedded within six disparate proteins suggests they may have shared genetic information at some point in their evolution. Only a segment of each protein is shown; it corresponds to a single identifiable domain. The similarities within the domain are unmistakable, even though some of the proteins greatly elsewhere in their structures. It appears that DNA ing the domain has been copied from gene to gene. The prote all recent products of evolution, found only in vertebrate an

Sequence and Structure. Their goal was to publish annually "all the sequences that could fit between a single pair of covers." It soon became apparent, however, that the covers would have to be very far apart, and computer tapes began to replace bound volumes as the working medium of the sequence comparer. Today any investigator can gain access to large sequence banks from a computer terminal.

About 10 years ago, working with a tape of data from the *Atlas,* I began to study the phylogeny of certain proteins. I was soon maintaining my own data bank, and whenever a new sequence was reported, I would enter it in the archive to see if it resembled anything already known. The number of matches was surprisingly large.

I should like to give an example of how this molecular paleontology works. In the late 1970's Staffan Magnusson and his co-workers at the University of Aarhus in Denmark determined the amino acid sequence of antithrombin III, a protein in the blood plasma of vertebrate animals. Antithrombin III neutralizes thrombin, a blood-clotting factor whose mode of action is that of a protease, or protein-cutting enzyme. At about the same time a second group reported the sequence of alpha-1 antitrypsin, another protease inhibitor in the blood plasma. The Danish group compared the two sequences and found they were identical at 120 of 390 sites, a homology of about 30 percent. It seemed obvious they had descended from a common ancestral protein.

Not long after, workers at the National Biomedical Research Foundation at Georgetown University entered into their computer the sequence of ovalbumin, a protein abundant in egg white. They found that it resembles antithrombin III and alpha-1 antitrypsin, again to the extent of about 30 percent. The discovery came as a surprise, because up to then no one had any idea what the function of ovalbumin might be. The possibility that it is a protease inhibitor now had to be considered.

In 1983 a Japanese group published the sequence of angiotensinogen, the precursor of a small peptide hormone that regulates blood pressure. Although the hormone itself is only 10 amino acids long, the precursor extends to about 400 units. When I compared the sequence of angiotensinogen with the sequences in my data bank, the search revealed a low-level resemblance to alpha-1 antitrypsin. The resemblance was one of those in the twilight zone, amounting to only a 20 percent identity, but a statistical analysis convinced me the two proteins are members of the same family. Since then corroborating observations have been made by others, and there is no doubt of the kinship.

Another Danish group has recently added a fifth branch to this unexpected tree of related proteins: it is a substance of unknown function found in barley seeds and called protein *Z.* Although protein *Z* is only half the size of the others (about 200 amino acids), it is clearly related to them. Indeed, the half size fits well with experimental findings that the other proteins in the family have two major domains.

The discovery of these five related proteins in diverse settings suggests two lessons. First, whether or not the 4,000 amino acid sequences known today represent a significant fraction of all proteins, a point has been reached where any newly determined sequence has a good chance of resembling one already on record. Second, certain large-scale arrangements of amino acids are so useful in biochemistry that they have been employed over and over again in different contexts. Often these functional units can be identified with the domains recognized in structural studies.

One of the most widely distributed domains was discovered in 1974 by Michael G. Rossmann and his colleagues at Purdue University. They noted from X-ray-diffraction maps that several enzymes had an important feature in common: even though the overall structures of the proteins were quite different, they all included a domain of about 70 amino acids with essentially the same folding pattern. The enzymes also differed greatly in function, but they had in common the ty to bind certain coenzymes, r ly nicotinamide adenine dinucl (NAD), flavin mononucleotide (F or adenosine monophosphate (A All these molecules include a n nucleotide within their structure ubiquitous domain in the enzyr the binding site for the mononu tides, and Rossmann named it the onucleotide fold.

The discovery led Rossmann bold hypothesis. The domain fou all these enzymes, he proposed, ghost of a primitive protein fron cellular times. Its ability to bind r otides was so important that i incorporated into the machine several of the prototype enzyme emerged in the first living system still recognizable today.

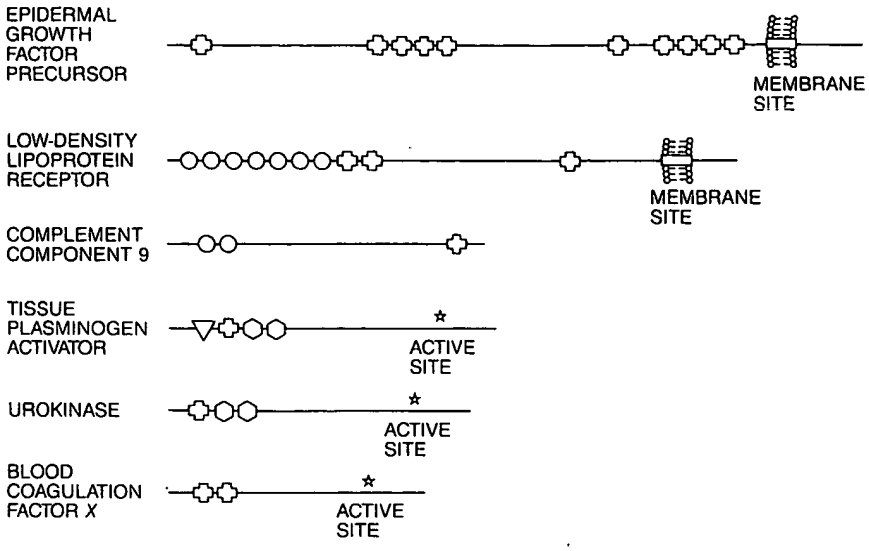
The model is an attractive o seems likely that the first functi proteins were small and that the portant capability was the bindi other molecules. If two small pr able to bind two different small cules were joined, the rudiments talysis could be initiated. Once ed, a succession of gene duplic could lead to an extended fam stable proteins. In the early stag loosely gathered proteins wou clumsy and inefficient. Opportu for improvement would be n ous, however, and the natural tion of mutant structures would inate events. Eventually the pr would be so artfully suited to function and the enzymes so eff that "natural rejection" of mutan ants would prevail.

Not long after Rossmann su ed that primitive proteins have been created by the fusion o ful domains, recombinant-DNA ies led to the startling discover eukaryotic genes are not contin They consist of segments that e part of a protein's structure (e separated by long stretches of no ing DNA (introns). In some cas introns were observed to fall at o the boundaries of a protein's don

This correspondence led Walte bert of Harvard University to pr

that exons are the genomic equivalent of the interchangeable protein parts hypothesized by Rossmann. In Gilbert's view, not only were the first proteins created by the assembly of stable domains but also evolution had maintained the genetic isolation of the domains over the course of several billion years. It is easy to see how this genomic organization might convey an adaptive advantage: the continuing reassembly of domains in new combinations would give rise to novel, and occasionally useful, proteins. Gilbert's ideas have been widely accepted, although there are also counterarguments. In many eukaryotic proteins introns fall at places other than obvious domain boundaries. Furthermore, prokaryotic genes have no introns at all; it is necessary to suppose they were eliminated in the interest of genomic economy.

Lately another remarkable instance of the dispersal of domains throughout a group of proteins has come to light. In this case the proteins are all recent products of evolution; they are found only in vertebrate animals that arose well within the past billion years. Moreover, the distribution of the domains among these proteins cannot readily be explained as a simple result of descent from a common ancestor. One domain is present in 18 copies scattered throughout six proteins. It seems clear these subunits have been passed freely from one protein to another and have been inserted wherever their functional activity is needed. For several of the proteins it has been shown that the DNA coding for the domains is precisely delimited by introns. In these cases there can be no question that the organization of the genome into exons and introns has been instrumental in the rearrangement of the mosaic gene products.



SHUFFLING OF MULTIPLE DOMAINS is evidence of the continual diffusion of genetic information in higher organisms. Five domains are represented by various geometric symbols, and their distribution is shown in six proteins. The domain whose sequence is given in the illustration on the opposite page is the one marked here by a cross. The distribution of the domains cannot readily be explained by assuming that all the domains in a given protein were inherited from the same ancestral gene; instead they seem to have spread from one protein to another by chromosomal rearrangements. In several cases boundaries between domains in the protein correspond to boundaries between exons and introns in the genome, which may have facilitated the shuffling of gene segments in the course of evolution.

Does this confirm Gilbert's hypothesis that exonic shuffling of domains has been a major feature of protein evolution from the earliest times? Although such shuffling is certainly going on now, I think it is a mistake to assume the same mechanism was at work in more primitive organisms. Introns can be dealt with in eukaryotic genes only because sophisticated splicing machinery ensures that the pieces of messenger RNA are properly translated into protein. It seems unlikely the same apparatus could have been present in the earliest life forms. Exon exchange in the mosaic vertebrate proteins is more likely a reenactment of ancient events, but in a totally modern guise.

Such variations on a theme are to be expected in a system as complex as the living cell, where a change in one molecule can affect thousands of others, including the very machinery responsible for synthesizing the first molecule. Just as proteins evolve, so do the mechanisms of protein evolution.